

EXHIBIT 28

Case No. 3:23-cv-03417-VC
Attorney's Eyes Only

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

RICHARD KADREY, an individual, et
al.

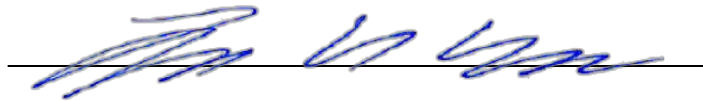
v.

META PLATFORMS, INC., a Delaware
corporation;
Defendant.

Case No. 3:23-cv-03417-VC

EXPERT REPORT OF LYLE UNGAR, PH.D.

Signed in Philadelphia, Pennsylvania on January 10, 2025



Lyle Ungar, PhD

responses.⁴² The network's weights are then iteratively adjusted to improve its predictions.⁴³ After training is complete, the network's weights have been trained to make accurate predictions on new input data not seen in training. Importantly, neural networks can consist of billions of weights, with each weight typically represented by a decimal value.⁴⁴ A sample of a neural network's weights (only 100 weights of billions total) is shown below.

Figure 7 - Sample Neural Network Weights

```
[ 1.2516975e-06, -1.7881393e-06, -4.3511391e-06, 8.0466270e-06,
  1.9073486e-06, -5.6028366e-06, 3.0994415e-06, 1.1920929e-06,
  -6.7949295e-06, -1.6689301e-06, -4.4703484e-06, -4.4107437e-06,
  -7.1525574e-07, -8.4638596e-06, 2.1457672e-06, 1.0251999e-05,
  -4.7683716e-07, -1.5497208e-06, 1.6093254e-06, 1.1324883e-06,
  2.8610229e-06, 9.4771385e-06, 3.3378601e-06, -2.8014183e-06,
  -1.2874603e-05, -2.8014183e-06, 5.6028366e-06, -1.1324883e-06,
  -3.3378601e-06, -2.9802322e-06, -2.3841858e-07, 1.4305115e-06,
  9.1791153e-06, 2.5629997e-06, 1.9669533e-06, 9.5367432e-07,
  -1.1086464e-05, -5.7220459e-06, 3.9339066e-06, -1.1026859e-05,
  7.2121620e-06, 1.8477440e-06, 4.5895576e-06, 2.2053719e-06,
  1.3113022e-06, -2.8610229e-06, -1.4841557e-05, -6.4373016e-06,
  2.6226044e-06, 6.6757202e-06, -2.6226044e-06, 8.3446503e-06,
  1.6093254e-06, -1.3053417e-05, 4.6491623e-06, 7.8082085e-06,
  -5.6028366e-06, -6.3180923e-06, -3.5762787e-07, 9.2387199e-06,
  5.3644180e-06, -3.9339066e-06, -2.4437904e-06, -4.6491623e-06,
  -6.9737434e-06, 1.9073486e-06, 4.0531158e-06, -2.9206276e-06,
  2.6822090e-06, 9.1791153e-06, -6.1988831e-06, 5.6624413e-06,
  1.4901161e-06, 8.1658363e-06, -3.6954880e-06, 7.0333481e-06,
  -1.6689301e-06, 2.8610229e-06, 8.4042549e-06, -5.1259995e-06,
  8.2254410e-06, -3.6954880e-06, 9.5367432e-06, 1.5497208e-06,
  -4.2915344e-06, 2.0265579e-06, -2.0265579e-06, -1.5497208e-06,
  -8.9406967e-06, -1.1920929e-07, 4.7683716e-06, 2.6226044e-06,
  6.9737434e-06, 9.5963478e-06, 3.4570694e-06, -7.6889992e-06,
  -1.9073486e-06, -5.0067902e-06, 6.4373016e-06, 5.3644180e-06, ... ]
```

Neural Network Weights: Weights in Neural Networks. Modern Neural Networks often include millions or billions of weights.

51. As the complexity and sophistication of deep learning models increases, the need for larger quantities of data is very important, particularly for estimating the weights accurately.⁴⁵ In training, models learn from the patterns shown in their training input data. To effectively train

⁴² "What Is Supervised Learning? | IBM," December 28, 2024, <https://www.ibm.com/think/topics/supervised-learning>.

⁴³ Deep Learning, Goodfellow, Bengio, Courville, p.106.

⁴⁴ Artificial Intelligence: A Modern Approach, Russel & Norvig, p.1211.

⁴⁵ Artificial Intelligence: A Modern Approach, Russel & Norvig, p.77-78.

a neural network, training data must be of significant *scale, variety, and quality*.⁴⁶ Later in this report, I discuss in more detail how these three qualities: scale, variety, and quality impact training data requirements for Large Language Models (LLMs).

3) Supervised, Unsupervised, and Self-Supervised Learning

52. Depending on the data that is available for training a machine learning model, various methods and structures exist by which models can learn from the data. This distinction in the available data divides machine learning techniques into supervised, unsupervised, and self-supervised learning.⁴⁷ As the names suggest, these methods relate to how machine learning models learn from their training data.

53. As discussed above, *supervised learning* trains a model to predict a specific target by allowing the model to learn from instances of input data having an associated target data value.⁴⁸ The Zillow model discussed above is an example of supervised learning, where the model learns to relate the sale price of a home with its square footage. In this sense, the training data can be thought of as a guide that is *supervising* the model's learning process—the model attempts to predict the sale price of a home and is guided by the pre-existing relationship between the square footage and sales prices that is available in the training dataset.

54. *Unsupervised learning*, on the other hand, does not contain the target data as part of the input data. Since the target variable is missing from the input data, the unsupervised machine learning models must understand relationships between various characteristics of the input data, and group them together in order to derive patterns from the relationship within the input data itself.⁴⁹ As one example of unsupervised learning, Spotify, a popular audio content streaming platform, employs unsupervised learning to power its suggestion for the “Discover Weekly” feature.⁵⁰ Spotify's recommender system tries to identify content that a user may

⁴⁶ Artificial Intelligence: A Modern Approach, Russel & Norvig, p.77-78.

⁴⁷ Deep Learning, Goodfellow, Bengio, Courville, pp.20-21.

⁴⁸ Probabilistic Machine Learning: An Introduction, Murphy, pp. 1-2.

⁴⁹ Probabilistic Machine Learning: An Introduction, Murphy, p.14.

⁵⁰ The Data School, “Machine Learning 101 and The Spotify Case,” The Data School Down Under (blog), April 2, 2023, <https://www.thedataschool.com.au/mipadmin/machine-learning-101-and-the-spotify-case/>.

data used by Llama 3 is textual data sourced from the web. For such data, Meta first identified repeating URLs within the dataset and kept only the most recent version of each URL found in the dataset. Hence, the text data from one website only appears once in the cleaned training dataset. Further, to deduplicate text that is seen across multiple webpages, Meta utilized “line-level deduplication,” aimed at removing any lines that appear more than 6 times among the webpages in the dataset. As a result, Meta removed very common, or boilerplate, lines of text (such as cookie warnings) from their webpage training data.³¹⁷

215. Additionally, for any documents such as books or scientific publications used in training Llama 3, Meta utilized a “MinHash” technique to identify documents that are approximately the same as other documents.^{318, 319} A MinHash algorithm attempts to calculate the approximate similarity between two sets of data, by comparing the shared values over the entire dataset and calculate a similarity score. By using this algorithm, Meta was able to identify near duplicate documents and remove documents with a high level of similarity.

216. Finally, Meta leveraged other LLMs to ensure that only high-quality data (i.e., data capturing useful semantic relationships, such as well cited academic publications) was included in the training dataset.³²⁰ In contrast, low quality data can contain biased inputs or introduce inaccurate information to the LLM, which can affect its accuracy and overall performance.³²¹ For each document used in the training data, the practitioners used classifiers developed using other LLMs to determine whether the document contains high-quality text data.

217. These steps for cleaning the training dataset were applied across different kinds of textual data, such as multilingual text, code, and reasoning text data. In addition to textual data, Meta

³¹⁷ Abhimanyu Dubey et al., “The Llama 3 Herd of Models” (arXiv, August 15, 2024), <http://arxiv.org/abs/2407.21783>.

³¹⁸ Esiobu Deposition 130:9-132-8.

³¹⁹ See 20241115/Fairspark_FullAllVersion+PulRequests/fairspark/utils/minhaslsh.py

³²⁰ Abhimanyu Dubey et al., “The Llama 3 Herd of Models” (arXiv, August 15, 2024), <http://arxiv.org/abs/2407.21783>.

³²¹ Max Lukichev, “Data Quality’s Role in Advancing Large Language Models,” Telmai, September 20, 2023, <https://www.telm.ai/blog/demystifying-data-qualitys-impact-on-large-language-models/>.

datasets are often shared publicly and freely with the broader NLP community. Practitioners commonly download shared datasets, assess their usefulness, incorporate them into their research and products, and describe their use cases in industry publications.

234. As LLMs learn to approximate broad patterns it encounters during training (as discussed in **Section IV.B.2.b**), practitioners remove low-quality datasets such as those with abundant misspellings, toxic language, or misinformation, from LLM training datasets.³⁵¹ Filtering low-quality data before pretraining can often be resource intensive and difficult to implement for large datasets, since it requires processing trillions of words of text and often reviewing data samples manually.³⁵² As a result, practitioners ultimately rely on trusted datasets that have been proven to contain high-quality data and have previously been used in the industry to successfully train similar models. Further, practitioners having access to the same datasets allows for the production of reproducible scientific results.

A. The Pile and Books3 are Widely Used in Research and the Industry

235. Certain popular public datasets are frequently used by NLP researchers and engineers. Books3 and The Pile, which I describe in **Section IV.B.2.c** above, are two such datasets. The Pile is a large, diverse text dataset assembled by EleutherAI, a non-profit AI research group,³⁵³ containing nearly 1 trillion words, and composed of 22 sub-datasets, including text from various sources such as encyclopedias, webpages, social media, and movie subtitles, among others³⁵⁴ Further, Books3 is one of the 22 subsets of The Pile dataset, containing the texts of thousands of books.³⁵⁵ Other, similar datasets compiled by NLP researchers and released publicly include Common Crawl, WebText, Colossal Cleaned Common Crawl (C4), Wikipedia, BooksCorpus, Project Gutenberg books published before 1919 (PG-19),

³⁵¹ Guilherme Penedo et al., “The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale” (arXiv, October 31, 2024), <https://doi.org/10.48550/arXiv.2406.17557>.

³⁵² Vasu Sharma et al., “Text Quality-Based Pruning for Efficient Training of Language Models” (arXiv, May 10, 2024), <https://doi.org/10.48550/arXiv.2405.01582>.

³⁵³ Leo Gao et al., “The Pile: An 800GB Dataset of Diverse Text for Language Modeling” (arXiv, December 31, 2020), <http://arxiv.org/abs/2101.00027>.

³⁵⁴ Leo Gao et al., “The Pile: An 800GB Dataset of Diverse Text for Language Modeling” (arXiv, December 31, 2020), <http://arxiv.org/abs/2101.00027>.

³⁵⁵ Leo Gao et al., “The Pile: An 800GB Dataset of Diverse Text for Language Modeling” (arXiv, December 31, 2020), <http://arxiv.org/abs/2101.00027>.

Responsible Open-science Open-collaboration Text Sources (ROOTS), and RedPajama-V2.³⁵⁶ As I highlight in the following paragraphs, The Pile and more specifically Books3 are not only popular among research practitioners, but also used ubiquitously by public and private companies, research organizations, and academic institutions for NLP tasks.

236. EleutherAI's original research publication introducing and describing the contents of The Pile has been cited over 600 times in scholarly journals and preprints,³⁵⁷ and over the six months the dataset was available for download from Hugging Face, it was downloaded more 68,000 times.³⁵⁸ The GitHub repository containing the dataset, and its preprocessing source code has been forked over 100 times.^{359,360} Further, The Pile has been referenced in seminal publications in top NLP and computer science journals and conferences, including the Journal of Machine Learning Research (JMLR), the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Learning Representations (ICLR), the Association for Computational Linguistics (ACL) annual meeting, and the American Journal of Computational Linguistics. Over 100 organizations, including technology companies (such as Google, Microsoft, and Apple), universities (such as Stanford, MIT, and Cornell), research institutions (such as the Allen Institute for AI and the Vector Institute for Artificial Intelligence) and other non-profit organizations have produced research publications utilizing The Pile and Books3, which collectively have been cited over 10,000 times according to Google Scholar.

³⁵⁶ Yang Liu et al., "Datasets for Large Language Models: A Comprehensive Survey" (arXiv, February 27, 2024), <http://arxiv.org/abs/2402.18041>.

³⁵⁷ Leo Gao et al., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" (arXiv, December 31, 2020), <http://arxiv.org/abs/2101.00027>.

³⁵⁸ Based on monthly counts available on Wayback Machine snapshots of the EleutherAI HuggingFace page, which can be found at "Wayback Machine," accessed on January 3, 2025. <https://web.archive.org/web/20230605134155/https://huggingface.co/datasets/EleutherAI/pile>.

³⁵⁹ "EleutherAI/the-Pile," Python (2020; repr., EleutherAI, November 12, 2024), <https://github.com/EleutherAI/the-pile>.

³⁶⁰ A fork is a new repository that shares code with the original repository. Essentially, forking on GitHub allows practitioners to use a repository, the code used to create The Pile in this case, as a baseline for their own developments in the future, and is a popular method to develop new applications using open-source code. Forking also enables developers to build on the underlying code from the forked repository to create new applications. See: "Fork a Repository," GitHub Docs, accessed September 28, 2024, <https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/working-with-forks/fork-a-repo>.

237. Usage of The Pile in scholarly work has advanced the state of NLP in distinct ways. Specifically, studies leveraging The Pile have created novel data filtration and weighting techniques, new LLM architecture optimizations to improve task performance on long texts, and general LLM pretraining performance improvements. The following paragraphs expand upon these research examples and describe several NLP advancements made possible by The Pile and Books3 datasets.
238. First, The Pile has served as a strong training dataset when studying data filtration techniques and their impacts on LLM training. For example, a 2023 publication from researchers at MIT and Cornell, in partnership with researchers from Google and OpenAI, examined the impact of various pretraining text toxicity and quality filters on trained LLM outputs, as well as assessing the relative impacts of LLM dataset components on output toxicity.³⁶¹ Using The Pile as an exemplar LLM training dataset, the researchers trained many small LLMs on various combinations of The Pile's sub-datasets, with various filters applied to remove unwanted text. The researchers proposed an optimal data mixture and filtration technique to maximize generalization and prevent toxic LLM outputs. A similar study by Stanford University and Google researchers analyzed various pretraining data mixtures to determine an optimal mix to speed up LLM pretraining.^{362,363} Again using The Pile as the exemplar LLM training dataset, their analysis resulted in a proposed data mixture that trains an LLM 2.6 times faster and improves benchmark performance accuracy by 6.5% by increasing the scale of web text, subtitles, and emails shown to the LLM during pretraining.
239. Second, The Pile, and Books3 in particular, provides a dataset of long texts that researchers may find useful for evaluating LLM architecture modifications that extend a model's context

³⁶¹ Shayne Longpre et al., "A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality & Toxicity" (arXiv, May 22, 2023), <https://doi.org/10.48550/arXiv.2305.13169>.

³⁶² A *pretraining data mixture* refers to the percentage of LLM training datasets from various domains. For example, one LLM's training text could be 80% data obtained from websites and 20% code from GitHub, while another's could be 65% website data, 15% books, and 20% academic papers.

³⁶³ Sang Michael Xie et al., "DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining" (arXiv, May 17, 2023), <https://doi.org/10.48550/arXiv.2305.10429>.

length.³⁶⁴ Models with longer context lengths are desirable because they are able to make predictions based on more text. For example, in 2023, researchers from Princeton University published a study demonstrating a method for LLMs to automatically summarize long texts before processing them, effectively extending context length indefinitely.³⁶⁵ Using Books3 as a reference set of long texts, the researchers designed an AutoCompressor³⁶⁶ architecture, trained it on two billion tokens from Books3, and evaluated its performance on text continuation and benchmark tasks. Experimenting with a context length of over 6,000 tokens, they demonstrated a 5.0% improvement on text continuation and a 6.4% improvement on a benchmark new article classification task. A separate effort by researchers at Hangzhou Dianzi University and University of Science and Technology of China augmented LLMs' positional embeddings³⁶⁷ to allow the model to understand token positions beyond its natural pretrained range.³⁶⁸ After training on a different dataset of long texts, the method is evaluated through testing on samples from Books3. In next-token prediction tests, the new method showed improvements over existing methods for context lengths between 16,000 and 128,000 tokens and produced usable results with over one million-token context lengths for the first time.³⁶⁹

240. Lastly, while in the examples above The Pile and Books3 contribute to data selection (before LLM pretraining), and LLM evaluation (after pretraining) techniques, The Pile and

³⁶⁴ As described in **Section IV.C.1, Context Length** is the maximum number of tokens that can be input into an LLM. Smaller LLMs often use between five hundred and one thousand tokens (approximately two pages of text), making it a limiting factor in many research LLM applications.

³⁶⁵ Alexis Chevalier et al., "Adapting Language Models to Compress Contexts" (arXiv, May 24, 2023), <https://doi.org/10.48550/arXiv.2305.14788>.

³⁶⁶ A variant of an LLM architecture that successively incorporates machine-readable summaries of earlier text into representations of later text.

³⁶⁷ I omitted *positional embeddings* from earlier descriptions of LLMs since they are complex and not necessary to gain a strong general understanding of how LLMs work. It suffices to say here that they are added to initial word embeddings (see **Section IV.B.1.a.ii**) just before an LLM's first transformer layer and are designed to give the model a sense of each token's positional placement within a sequence of text. They are needed because an LLM's architecture does not naturally provide any information about token order. Look at the attention diagram in **Figure 16**, for example—though some words are next to each other, each is compared equally with all other words in the sequence without regard for their respective positions. Positional embeddings give words that are close to each other a slightly stronger default connection.

³⁶⁸ Yiran Ding et al., "LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens" (arXiv, February 21, 2024), <https://doi.org/10.48550/arXiv.2402.13753>.

³⁶⁹ Yiran Ding et al., "LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens" (arXiv, February 21, 2024), <https://doi.org/10.48550/arXiv.2402.13753>.

Books3 have also been used extensively by researchers and companies to train LLMs. Researchers often train small LLMs on The Pile to evaluate a new technique's effectiveness. For example, researchers at the Chinese Academy of Sciences recently validated a new method for rewarding LLMs for generalization during training by implementing their method and then training multiple small LLMs on The Pile.³⁷⁰ Their method produced a 1.6% average improvement across eight benchmark tasks. Beyond academic research models, large institutions and companies often use The Pile for LLM training. NVIDIA's Megatron,³⁷¹ Stanford University's BioMedLM³⁷², and Apple's OpenELM³⁷³ were all trained on The Pile or large subsets thereof, and each presented innovations on standard LLM architecture and pretraining practices. The Pile is used so pervasively for LLM training that nearly every open-source LLM released between 2021 and 2023 included The Pile in its training data, either in its entirety or in large part.³⁷⁴

241. The publications highlighted above demonstrate the impact and widespread use of The Pile and Books3 datasets by practitioners across universities, research, and technology organizations, among others. In addition to their use pretraining large language models, the datasets are frequently used to evaluate new LLM architectures, assess model performance on new tasks, and optimize LLM training.

B. The Library Genesis Dataset is Compositionally Similar to Books3, though Larger, Older, and Not as Applicable for Many Smaller LLMs

242. Finally, another large book dataset is the Library Genesis (Libgen) dataset. Libgen contains scholarly articles, scientific, and fiction books, as well as text content from disparate

³⁷⁰ Zhenpeng Su et al., "MiLe Loss: A New Loss for Mitigating the Bias of Learning Difficulties in Generative Language Models" (arXiv, March 28, 2024), <https://doi.org/10.48550/arXiv.2310.19531>.

³⁷¹ "Nvidia/Nemo-Megatron-Gpt-20B · Hugging Face," May 3, 2023, <https://huggingface.co/nvidia/nemo-megatron-gpt-20B>.

³⁷² "Stanford-Crfm/BioMedLM · Hugging Face," January 26, 2024, <https://huggingface.co/stanford-crfm/BioMedLM>.

³⁷³ Sachin Mehta et al., "OpenELM: An Efficient Language Model Family with Open Training and Inference Framework" (arXiv, May 1, 2024), <https://doi.org/10.48550/arXiv.2404.14619>.

³⁷⁴ See Eugene Yan, "Eugeneyan/Open-Llms," January 6, 2025, <https://github.com/eugeneyan/open-llms> for a list of open-source LLMs. Every listed LLM released in 2021 or 2022 uses at least substantial portions directly from The Pile, with the dataset having been used by more than 400 publications since release. See: "Papers with Code - The Pile Dataset," accessed January 9, 2025, <https://paperswithcode.com/dataset/the-pile>.